Rapport final du projet jeunes chercheurs du GDR ISIS

BISOU
(BIStochastic Optimization and application to mUltiple kernel learning)

Sandrine Anthoine[1], Liva Ralaivola[2], Marie Szafranski[3], Matthieu Kowalski[4], Thomas Peel[1,2]

[1]LATP, CNRS, Aix-Marseille Université, 39, rue F. Joliot Curie, 13013 Marseille, France

[2]LIF, CNRS, Aix-Marseille Université, 39, rue F. Joliot Curie, 13013 Marseille, France

[3]IBISC, Université d'Évry Val d'Essonne, Boulevard François Mitterand, 91025 Evry cedex, France

[4]L2S, CNRS, Supélec, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France

anthoine@cmi.univ-mrs.fr, ralaivola@lif.univ-mrs.fr,szafranski@ibisc.fr,kowalski@lss.supelec.fr,peel@lif.univ-mrs.fr

Résumé – Le projet BISOU - *BIStochastic Optimization and mUltiple kernel learning* - avait pour but de développer des algorithmes stochastiques pour l'optimisation convexe (non différentiable) en grande dimension et traitant un large volume de données. Deux algorithmes itératifs et aléatoires ont été conçus, l'un sélectionnant à chaque étape un sous-ensemble de données et l'autre un sous-ensemble de dimensions. Tous deux permettent de traiter des données de grandes dimensions, le premier a été appliqué à un problème de traitement de signal Magnéto-Encéphalographique (MEG), le second à un problème d'apprentissage par régression.

Thèmes du GDR-ISIS concernés: Thèmes A (Méthodes et modèles en traitement de signal) et B (Image et vision).

1 Description du projet

1 Contexte scientifique

Dans ce projet, nous nous intéressons au problème d'optimisation efficace dans le cadre de larges masses de données. Nous partons de la formalisation suivante, commune à des problèmes en apprentissage statistique et en traitement de signal :

$$\min_{w\in\mathbb{R}^d}\left[J(w) = J_{reg}(w) + \frac{\lambda}{N}\sum_{i=1}^{N} l_i(w, z_i)\right], \quad (1)$$

où les $\{z_i\}_{i=1}^N$ sont les données, $l_i(w, z_i)$ est un terme d'attache à la donnée z_i et J_{reg} est un terme de régularisation incorporant des *a priori* sur la forme de w. Nous considérons la situation où la dimension des données d et/ou le nombre de données N peuvent être grands et où les différents termes (régularisation et attache aux données) sont convexes mais non nécessairement différentiables. Les problèmes à traiter sont les suivants : 1) les difficultés liées à la non-différentiabilité de la fonctionnelle à minimiser et 2) les difficultés liées au stockage et à la rapidité des calculs engendrés par la grande masse de données et leur dimension.

Nombreux travaux (plus ou moins récents) en optimisation convexe permettent de traiter la non-différentiabilité. Certains, comme les schémas itératifs multi-pas qui font usage à chaque étape des gradients calculés aux itérations précédentes permettent même d'obtenir des vitesses de convergence optimales [1, 7]. Toutefois ces schémas d'optimisation ne sont pas très adaptés au traitement de données de grandes dimensions ou de large masse de données. Pour cela, des algorithmes aléatoires visant à réduire soit

le nombre des donnés soit leur dimensionalité ont été développés. Des algorithmes de sélection des données ont été développés dans le cadre non-différentiable, comme par exemple [11, 2], mais dont on ne contrôle pas les vitesses de convergence. Dans le cadre de sélection aléatoire des coordonnées, une méthode multi-pas (à convergence optimale) est proposée par Nesterov dans [8] pour le cas où les termes de régularisation et d'attache aux données sont tous deux différentiables.

2 Projet initial

D'après nos recherches bibliographiques, aucun algorithme bi-stochastique procédant à la fois à la sélection des coordonnées et des données n'a encore été mis en place. Le projet BISOU avait donc comme objectif initial de proposer et étudier des algorithmes bi-stochastiques théoriquement fondés qui combinent à la fois la sélection de données et la sélection de coordonnées dans le cadre de schémas itératifs de minimisation de fonctionnelles du type de l'Eq. (1).

Pour cela, nous proposions de développer des algorithmes stochastiques de la forme suivante :

Algorithme 1.
- *Initialisation des variables.*
- *À chaque itération k :*

Sélection des données *Choix aléatoire d'un sous-ensemble* A_k *de* $\{1, .., N\}$.

Sélection des coordonnées *Choix aléatoire d'un sous-ensemble* B_k *de* $\{1, .., d\}$.

Mise à jour

- $h^k = \underset{h \in \mathbb{R}^d_{|B_k}}{\operatorname{argmin}} J^k_{reg}(w^k+h) + \frac{\lambda}{|A_k|}\sum_{i \in A_k} l^k_i(w^k+h, z_i)$,
- $w^{k+1} = w^k + h^k$.

où $\mathbb{R}^d_{|B_k}$ *est le sous-espace de* $\mathbb{R}^d$ *défini par les coordonnées de* B_k, J^k_{reg} *et* $l^k_i(w, z_i)$ *sont des approximations régularisées des termes de la fonctionnelle J (tenant compte par exemple du dernier itéré* w^k*).*

L'application visée était le problème de l'apprentissage multi-noyaux, suite aux travaux entrepris par les équipes du L2S, d'IBISC et du LIF [4], partenaires du projet.

Les verrous principaux du projet étaient :

L'efficacité. Pour que l'algorithme 1 puisse gérer de grandes masses de données et/ou des données de grandes dimensions, il est indispensable que les fonctionnelles de remplacement J^k_{reg} et $l^k_i(w, z_i)$ permettent des calculs rapides, notamment lors de l'évaluation de leur (sous)-gradients.

D'autre part la stratégie de choix des sous-ensembles aléatoires doit elle-aussi être rapide, tout en garantissant un bon comportement de l'algorithme (notamment sa convergence).

Le bien-fondé théorique. Un objectif de ce projet était de proposer des schémas aléatoires bien contrôlés, notamment par une analyse mathématique qui garantit leur convergence et si possible même l'étude de leur vitesse de convergence.

3 Partenaires

Le projet a réuni les équipes de traitement du signal du LATP et L2S, apportant leurs connaissances en optimisation non-convexe, et les équipes d'apprentissage statistique d'IBISC et du LIF contribuant sur les aspects stochastiques et l'apprentissage multi-noyaux.

2 Travaux réalisés

En premier lieu, une recherche bibliographique nous a permis de découvrir des méthodes sélection, comme celles de type *active set* et de dégager trois pistes de recherche pour développer tout d'abord des algorithmes stochastiques simples (i.e. pour lesquels soit les coordonnées soit les exemples sont sélectionnés).

La première piste a fait l'objet d'une collaboration entres les participants marseillais (LIF et LATP), P. Machart (doctorant au LIF et LSIS) et H. Glotin son co-directeur de thèse. Le travail a porté sur un algorithme stochastique d'approximation de rang faible dans le cadre de l'apprentissage multi-noyaux et est décrit dans la section 1.

La seconde piste est celle de l'analyse et la stochastisation des méthodes de type *active set*. Elle implique M. Kowalski (L2S) et S. Anthoine (LATP) et est faite en collaboration avec P. Weiss (MCF à l'INSA de Toulouse) et A. Gramfort (post-doctorant à Harvard, Etats-Unis). Ces travaux sont décrits dans la section 2.

Enfin la troisième piste envisagée était celle de l'adaptation des méthodes de descente de coordonnées proposée par Y. Nesterov [8] au cas de fonctions non-différentiables. En effet le travail de Y. Nesterov sur la descente par sélection de coordonnées n'était applicable qu'au cas de

fonctionnelles différentiables. Nos travaux dans ce cadre ont été devancés par ceux de M. Takáč [9] présentés à la conférence SPARS 2011.

1 Régression par apprentissage stochastique de noyaux de rang faible

Dans ce travail, nous avons développé un algorithme itératif avec une sélection aléatoire des coordonnées à chaque itération pour le problème de régression "Ridge" à noyau décrit ci-dessous. Ce problème implique la manipulation de la matrice du noyau K de grande dimension. Pour réduire le coût calcul des mises à jours de la solution, on peut sélectionner les coordonnées à traiter ce qui revient à sélectionner des noyaux de rang 1 pour approximer K, ce que nous faisons ici aléatoirement. La méthode s'appuie sur les travaux de Nesterov [8] pour ce qui de l'optimisation coordonnées par coordonnées. Une coordonnée correspondant à un noyau de rang 1, les calculs de mises à jour des noyaux de faible rang doivent eux aussi être efficace, ce qui est rendu possible par l'utilisation de formules de Woodbury [14].

1.1 *Kernel Ridge Regression* classique

Soit l'ensemble d'apprentissage $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ où y_i est l'étiquette de l'exemple $\boldsymbol{x}_i$. Les exemples sont plongés via l'application ϕ dans un espace de Hilbert à noyau reproduisant $\mathcal{H}$ associé au noyau k (ainsi $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$).

On cherche une fonction de régression de la forme $f(\boldsymbol{x}) = \langle \phi(\boldsymbol{x}), \boldsymbol{w} \rangle_{\mathcal{H}}$ qui agit dans $\mathcal{H}$ en minimisant la fonctionnelle suivante :

$$\min_{\boldsymbol{w}} \lambda \|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} (y_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i) \rangle)^2, \quad (2)$$

La solution est $\boldsymbol{w} = \frac{1}{2\lambda} \sum_{i=1}^n \alpha_i^* \phi(\boldsymbol{x}_i)$ i.e. la fonction de régression s'écrit :

$$f(\boldsymbol{x}) = \frac{1}{2\lambda} \sum_{i=1}^{n} \alpha_i^* k(\boldsymbol{x}_i, \boldsymbol{x}), \text{ avec}$$
$$\boldsymbol{\alpha}^* = 2(I + \tfrac{1}{\lambda} K)^{-1} \boldsymbol{y} \quad (3)$$

1.2 Problème posé

La matrice de Gram K est la matrice des $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. L'inversion d'une matrice de telle taille $(n \times n)$ est difficile, nous la remplaçons par une combinaison conique de noyaux de rang 1 :

$$\tilde{K}(\boldsymbol{\mu}) = \sum_{m \in \mathcal{S}} \mu_m \tilde{K}_m \quad \text{avec} \quad \mu_m \geq 0, \text{ et}$$

$$\tilde{K}_m = \boldsymbol{c}_m \boldsymbol{c}_m^T,$$

$\boldsymbol{c}_m$ représentant la colonne m de K. Les $\tilde{K}_m$ sont appelées approximation de Nyström de rang 1 la matrice de Gram K [13].

Le problème consiste alors à trouver la "meilleure" combinaison conique $\boldsymbol{\mu}$, de manière à minimiser (2) avec $\tilde{K}(\boldsymbol{\mu})$ à la place de K. On régularise le problème en $\boldsymbol{\mu}$ par un a-priori de parcimonie et on aboutit au problème d'optimisation *convexe* et *dérivable* :

$$\min_{\boldsymbol{\mu} \geq 0} \{F(\boldsymbol{\mu}) := \boldsymbol{y}^T (I + \tfrac{1}{\lambda} \tilde{K}(\boldsymbol{\mu}))^{-1} \boldsymbol{y} + \nu \sum_m \mu_m\}. \quad (4)$$

1.3 Résolution

Bien que (4) soit convexe et différentiable, la minimisation n'est pas simple. On doit en effet recourir à une méthode itérative, qui peut être coûteuse à cause de la grande dimensionalité de $\boldsymbol{\mu}$ et des inversions de matrice $(I + \frac{1}{\lambda} \tilde{K}(\boldsymbol{\mu}))^{-1}$ impliquées.

Pour résoudre le problème de manière efficace et avec des garanties de convergences, de notre travail repose sur deux points clés :

- Nous avons mis au point un algorithme de descente de Newton aléatoire, où une coordonnée de $\boldsymbol{\mu}$ est choisie au hasard à chaque étape. En s'appuyant sur les travaux de [8], nous prouvons la convergence de cet algorithme aléatoire.
- Nous utilisons les formules de Woodbury pour l'inversion des matrices de rang 1 composant l'approximation de Nyström $\tilde{K}(\boldsymbol{\mu})$. Ceci permet des faire ces mises à jour de manière efficace.

Ces travaux ont été appliqués à des données tests type en régression et présentés dans les conférences ICML 2011 et CaP 2011 [6].

2 Accélération des algorithmes de seuillages itératifs par une stratégie de type "Active Set"

Ces travaux ont pour but de présenter une version aléatoire des algorithmes de sélection de variables par "Active Sets" tels celui de Roth et Fisher [10]. L'accent est mis sur la preuve de convergence d'un tel algorithme aléatoire et l'application au problème inverse de reconstruction des sources cérébrales en Magnéto-Encéphalographie (MEG).

Dans l'agorithme de Roth et Fisher, on utilise à chaque itération la solution exacte de la min-

imisation de la fonctionnelle restreinte aux variables de l'"Active Set". L'algorithme proposé remplace cette solution exacte par une solution approximée obtenue par quelques itérations de l'algorithme proximal correspondant à l'a-priori de parcimonie imposé [3].

La preuve de convergence est basée sur le fait que l'algorithme peut être vu comme une instance particulière d'un algorithme de descente de gradient coordonnée par coordonnée utilisant la règle *Gauss-Southwell-r* [12].

Les expériences effectuées prouvent que la méthode se révéle plus rapide que des accélérations de méthodes proximales type FISTA [1] pour certains problèmes inverses sous-déterminés utilisant des a-priori de parcimonie. C'est le cas en particulier pour le problème de reconstruction des sources cérébrales en MEG, lorsqu'on utilise un apriori de type "group-Lasso" (norme ℓ_{21}) spatio-temporel.

Ces travaux ont été présentés au "4th International Workshop on Optimization for Machine Learning" afférent à la conférence NIPS 2011 [5].

3 Bilan

Les partenaires du projet ont pu grâce à l'allocation du GDR ISIS se rencontrer pour travailler ensemble tous les 6 mois. La collaboration a été fructueuse, deux algorithmes stochastiques ont été développés et étudiés, l'un échantillonnant aléatoirement les données, l'autre les coordonnées. Les preuves de convergences ont été fournies. L'étude des vitesses de convergence fera l'objet de futurs travaux. Les applications ont touché l'apprentissage multi-noyaux et le traitement de signal. L'objectif final du projet initial - à savoir établir des algorithmes bistochastiques sélectionnant à la fois des variables et des coordonnées - n'a pas été atteint, mais les partenaires pensent que cela reste un but atteignable bien que plus ardu qu'imaginé au dépôt du projet. La collaboration entre les partenaires continuera au délà du projet GDR ISIS par le dépôt de projet type PEPS ou ANR.

Le bilan financier détaillé se trouve dans le document joint à ce rapport.

Références

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. *ICASSP*, pages 693–696, 2009.

[2] H. Chonghai, J. T. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, pages 781–789, 2009.

[3] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4) :1168–1200, November 2005.

[4] M. Kowalski, M. Szafranski, and L. Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *ICML*, pages 545–552, 2009.

[5] M. M. Kowalski, P. Weiss, A. Gramfort, and S. Anthoine. Accelerating ista. with an active set strategy. In *Workshop NIPS Opt2011*, 2011.

[6] P. Machart, T. Peel, S. Anthoine, L. Ralaivola, and H. Glotin. Stochastic low-rank kernel learning for regression. In *Proc. of the 28th International Conference on Machine Learning*, June 2011.

[7] Y. Nesterov. Gradient methods for minimizing composite objective function. Core discussion papers, Sept. 2007.

[8] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. Core discussion papers, 2010.

[9] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Submitted to Mathematical Programming*, 2011.

[10] V. Roth and B. Fischer. The group-lasso for generalized linear models : uniqueness of solutions and efficient algorithms. In *Proc. of ICML '08*, pages 848–855, 2008.

[11] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos : Primal estimated subgradient solver for svm. In *ICML*, 2007.

[12] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming B*, 117 :387–423, 2009.

[13] C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, 2001.

[14] M. A. Woodbury. Inverting modified matrices. Technical report, Statistical Research Group, Princeton University, 1950.